X-643-67-239

NASA TM X-55799

OPTIMAL COMPUTING FORMS FOR THE TWO-BODY C AND S SERIES

C. R. HERRON
E. R. LANCASTER
W. R. TREBILCOCK

FACILITY FORM 602

N67-28798 (ACCESSION NUMBER)

(THRU)

30 (PAGES)

1 (CODE)

TM-X-55799 (NASA CR OR TMX OR AD NUMBER)

19 (CATEGORY)

GPO PRICE $ ______

CFSTI PRICE(S) $ ______

Ha' .copy (HC) 3.00

Microfiche (MF) .65

ff 653 July 65

MAY 1967

GODDARD SPACE FLIGHT CENTER
GREENBELT, MARYLAND

X-643-67-239

OPTIMAL COMPUTING FORMS FOR THE TWO-BODY C AND S SERIES

C. R. Herron
E. R. Lancaster
W. R. Trebilcock

May 1967

GODDARD SPACE FLIGHT CENTER
Greenbelt, Maryland

OPTIMAL COMPUTING FORMS FOR THE TWO-BODY C AND S SERIES

INTRODUCTION

The classical solutions of the two-body problem separate naturally into the three cases of elliptic, parabolic, and hyperbolic motion, the mathematics being considerably different for each case. A unified formulation is possible, valid for all three cases, if certain transcendental functions, which we call the C and S functions, are introduced.

The unified formulation is fully developed by Battin [1] and will not concern us here. The purpose of this paper is the presentation of approximations for the C and S functions and their derivative functions which reduce significantly the computation times required for their evaluation when compared to that required by Taylor series expansions.

THE C AND S FUNCTIONS

The C and S functions are defined by

$$S(x) = \left(x^{1/2} - \sin x^{1/2}\right)/x^{3/2}, \qquad x > 0 \tag{1}$$

$$= \left[\sinh(-x)^{1/2} - (-x)^{1/2}\right]/(-x)^{3/2}, \qquad x < 0 \tag{2}$$

$$C(x) = \left(1 - \cos x^{1/2}\right)/x, \qquad x > 0 \tag{3}$$

$$= \left[1 - \cosh(-x)^{1/2}\right]/x, \qquad x < 0 \tag{4}$$

Since these functions are indeterminate for x = 0 and present accuracy problems when evaluated in the neighborhood of x = 0, it is natural to replace the above forms by the following series, convergent for all values of x:

$$S(x) = \sum_{i=0}^{\infty} \frac{(-x)^i}{(2i+3)!} \quad , \qquad (5)$$

$$C(x) = \sum_{i=0}^{\infty} \frac{(-x)^i}{(2i+2)!} \quad . \qquad (6)$$

For large values of x, the convergence of these series will be slow. It is then convenient to use the following reduction formulas, easily derived from the definitions (1) through (4):

$$A(x) = 1 - xS(x) \quad , \qquad (7)$$

$$2C(4x) = \left[A(x)\right]^2 \quad , \qquad (8)$$

$$4S(4x) = S(x) + A(x)\,C(x) \quad . \qquad (9)$$

THE C' AND S' FUNCTIONS

The derivatives $S'(x)$ and $C'(x)$ are needed for certain problems of orbit determination, guidance, and optimization. From (1) through (4) we obtain

$$S'(x) = \left[C(x) - 3S(x)\right]/2x \quad ,$$

$$C'(x) = \left[A(x) - 2C(x)\right]/2x \quad .$$

These forms suffer accuracy problems in the neighborhood of $x = 0$, again forcing us to series representations. Differentiating (5) and (6), we have

$$S'(x) = \sum_{i=1}^{\infty} \frac{i(-x)^i}{(2i+3)!} \quad , \qquad (10)$$

$$C'(x) = \sum_{i=1}^{\infty} \frac{i(-x)^i}{(2i+2)!} \quad , \qquad (11)$$

convergent for all values of x.

For large values of x, the following reduction formulas (obtained by differentiating (7), (8), and (9)) are useful.

$$B(x) = S(x) + xS'(x) \quad , \qquad (12)$$

$$C'(4x) = -A(x)B(x) \quad , \qquad (13)$$

$$4S'(4x) = S'(x) + A(x)C'(x) - B(x)C(x) \quad . \qquad (14)$$

THE FIKE-KNUTH ALGORITHM

Our first step in obtaining economical computing forms for (5), (6), (10), and (11) was the construction of sixth degree polynomial approximations on various intervals in the sense of Chebyshev. In other words, these polynomials minimize the magnitude of the maximum error on the interval. The program to accomplish this was written by the third author, based on ideas of Stoer [2]. The coefficients of these polynomials are given in Numerical Results.

Assume the approximating polynomial has the form

$$P(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6 \quad . \qquad (15)$$

The evaluation of (15) by the usual method of nested multiplication requires 6 multiplications and 6 additions. However, using recently developed polynomial evaluation methods [3, 4], (15) can be evaluated with 4 multiplications and 7 additions. The form and parameters for the algorithm, as it applies to our functions, are given in Numerical Results.

In the following description of the algorithm, a_6 is assumed to be positive. If a_6 is negative, a minor change is necessary.

Fike's modification of Knuth's method begins with a conversion: let $\mu = \sqrt[6]{a_6}$, and let $c_k = a_k/\mu^k$ for $k = 0, 1, \ldots, 5$. Then compute

$$p = \frac{1}{2}(c_5 - 1) \qquad D'' = c_2 - pC'$$

$$B' = c_4 - p(p+1) \qquad E' = 2D' - B' + 1$$

$$C' = c_3 - pB' \qquad E'' = 2D'' - B'D' - C'$$

$$D' = p - B' \qquad E''' = c_1 - B'D''$$

Find a real root q of the cubic equation*

$$2q^3 + E'q^2 + E''q + E''' = 0 \tag{16}$$

and compute

$$A = \frac{1}{2}B' - q$$

$$C = p - 2A$$

$$B = q - 2AC - A^2$$

$$D = C' - q(1 + D') - q^2 - D'' - A^2(1+C) - BC$$

$$E = q^2 + qD' + D'' - (A^2 + B)C$$

$$F = c_0 - (q^2 + qD' + D'')\left[C' - q(1+D') - q^2 - D''\right].$$

*See Appendix I.

Then our polynomial can be evaluated as follows:

$$q_1 = \mu x$$

$$q_2 = (q_1 + A)^2$$

$$q_3 = (q_2 + B)(q_1 + C)$$

$$P(x) = (q_2 + q_3 + D)(q_3 + E) + F$$

In case $a_6 < 0$, let $T(x) = -P(x)$ and perform all the steps above, except the last, for $T(x)$. The last step should be

$$P(x) = -[T(x)] = (q_2 + q_3 + D)(-q_3 - E) - F \ .$$

If the machine being used has a "load negative" feature which is equivalent in execution time to "load positive", and if subtraction is likewise equivalent to addition, then this modification is equivalent to the original.

As Fike points out, his method is a slight variation of that of Knuth [4], and since Knuth's method was inspired by Motzkin [5], the three types bear a strong family resemblance. Each begins with a polynomial in form (15) with $a_6 = 1$, and solves for the parameters in the final evaluation scheme by expanding the scheme into a sixth degree polynomial and equating its coefficients with those of form (15). To admit treatment of the general polynomial of degree six, however, some transformation must be made so that $a_6 = 1$. The most straightforward way is

$$Q(x) = P(x)/a_6$$

and then applying any of the three methods to $Q(x)$, adding an extra step at the last in multiplying the result by a_6. Fike specifies a different sort of transformation; his may be thought of as converting form (15) into

$$R(x) = x^6 + \frac{a_5}{\mu^5} x^5 + \frac{a_4}{\mu^4} x^4 + \frac{a_3}{\mu^3} x^3 + \frac{a_2}{\mu^2} x^2 + \frac{a_1}{\mu} x + a_0 \ .$$

Again, any of the three methods apply to $R(x)$ and values of $P(x)$ are obtained by using μx in the scheme for $R(x)$, since $R(\mu x) = P(x)$.

This transformation, though a bit more complicated, is admirably suited to our particular problem. The type of polynomial with which we are dealing has the not uncommon characteristic that

$$|a_6| < |a_5| < \ldots < |a_0|$$

and, in addition, $|a_6|$ is very small. For example, suppose the coefficients of form (15) are

$$a_6 = +.11 \times 10^{-10}$$

$$a_5 = -.21 \times 10^{-8}$$

$$a_4 = +.28 \times 10^{-6}$$

$$a_3 = -.25 \times 10^{-4}$$

$$a_2 = +.14 \times 10^{-2}$$

$$a_1 = -.42 \times 10^{-1}$$

$$a_0 = +.50$$

If we use the division transformation, the coefficients b_i of $Q(x)$ are

$$b_6 = +1.0$$

$$b_5 = -.18 \times 10^{3}$$

$$b_4 = +.24 \times 10^{5}$$

$$b_3 = -.22 \times 10^{7}$$

$$b_2 = +.12 \times 10^9$$

$$b_1 = -.36 \times 10^{10}$$

$$b_0 = +.44 \times 10^{11}$$

Here the errors in the numbers a_i have become greatly magnified; worse yet, the arithmetic of parameter production using the large numbers b_i is likely to suffer the effects of large error propagation. In contrast, Fike's transformation gives us

$$c_6 = 1$$

$$c_5 = -.27 \times 10^1$$

$$c_4 = +.54 \times 10^1$$

$$c_3 = -.73 \times 10^1$$

$$c_2 = +.62 \times 10^1$$

$$c_1 = -.28 \times 10^1$$

$$c_0 = .50$$

These numbers of manageable size lend themselves very well to whichever scheme we choose. For comparison, the two transformations above were evaluated by the Knuth algorithm for 40 points over the interval $[-1, +1]$, and the differences between these values and the true values of the polynomial were obtained. For the division transformation, the absolute value of the maximum error was $.92 \times 10^{-12}$; for the Fike transformation, this was $.16 \times 10^{-14}$, a reduction by a factor of more than 500. Several other test cases were run, with results which apparently verify the conclusion that the Fike transformation used on this type of polynomial has a very definite advantage. There are, of course, other transformations which produce polynomials in which $a_6 = 1$. In general, one should use the transformation which keeps the coefficients of the transformed polynomial as small as possible.

NUMERICAL RESULTS

The four approximation polynomials were generated for each of the intervals $[-1, +1]$, $[-2, +2]$, $[-4, +4]$, $[-16, +16]$, converted to form (15) and parameters for the Fike evaluation scheme were obtained. In each case, the values given by the final scheme were tested against "true" values of the original function for all multiples of .002 in the interval concerned. The "true" values came from expanding the power series of the function (1) for enough terms to guarantee that the relative error from truncation would be less than 10^{-15}. The following tables exhibit, for each of the sixteen functions considered, the coefficients a_i for form (15), the parameters A, B, C, D, E, and F for the Fike scheme, and the maximum absolute errors for both methods. For comparison, a degree 4 approximation polynomial was evaluated by both methods for the functions $C(x)$ and $S(x)$ on the interval $[-1, +1]$, and the results are presented here also.

$$C(X) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + a_6 x^6 \quad \text{on } [-h, h]$$

$h = 1$

$a_0 = +0.4999999999999998 \times 10^{0}$
$a_1 = -0.4166666666667176 \times 10^{-1}$
$a_2 = +o.1388888888888999 \times 10^{-2}$
$a_3 = -0.2480158725995993 \times 10^{-4}$
$a_4 = +0.2755731917059028 \times 10^{-6}$
$a_5 = -0.2087759200397967 \times 10^{-8}$
$a_6 = +0.1147134108311665 \times 10^{-10}$

$M_a = 0.76327833 \times 10^{-15}$

$h = 2$

$a_0 = +0.4999999999999993 \times 10^{0}$
$a_1 = -0.4166666666700118 \times 10^{-1}$
$a_2 = +0.1388888888892785 \times 10^{-2}$
$a_3 = -0.2480158663241807 \times 10^{-4}$
$a_4 = +0.2755731881710992 \times 10^{-6}$
$a_5 = -0.2088010277268315 \times 10^{-8}$
$a_6 = +0.1147215380312168 \times 10^{-10}$

$M_a = 0.95645714 \times 10^{-13}$

$h = 4$

$a_0 = +0.4999999999998401 \times 10^{0}$
$a_1 = -0.4166666668808485 \times 10^{-1}$
$a_2 = +0.1388888889138842 \times 10^{-2}$
$a_3 = -0.2480157659289839 \times 10^{-4}$
$a_4 = +0.2755731272513377 \times 10^{-6}$
$a_5 = -0.2089014196095935 \times 10^{-8}$
$a_6 = +0.1147636934013430 \times 10^{-10}$

$M_a = 0.12239224 \times 10^{-10}$

$h = 16$

$a_0 = +0.4999999894793170 \times 10^{0}$
$a_1 = -0.4166675473500692 \times 10^{-1}$
$a_2 = +0.1388889916034137 \times 10^{-2}$
$a_3 = -0.2479883633119184 \times 10^{-4}$
$a_4 = +0.2755565077419917 \times 10^{-6}$
$a_5 = -0.2109148028487573 \times 10^{-8}$
$a_6 = +0.1156091702389399 \times 10^{-10}$

$M_a = 0.20159773 \times 10^{-6}$

$$q_1 = \sqrt[6]{a_6}\ x$$
$$q_2 = (q_1 + A)^2$$
$$q_3 = (q_2 + B)(q_1 + C)$$
$$C(X) = (q_2 + q_3 + D)(+ q_3 + E) + F \quad \text{on } [-h, h]$$

$h = 1$

$A = +0.4513408582627891 \times 10^{0}$
$B = +0.3744865190483202 \times 10^{1}$
$C = -0.2769272800754423 \times 10^{1}$
$D = +0.9433565393074166 \times 10^{1}$
$E = +0.1055413968372178 \times 10^{2}$
$F = +0.6288190624578802 \times 10^{-2}$

$M = 0.5828670879282069 \times 10^{-14}$

$h = 2$

$A = +0.4512008438957284 \times 10^{0}$
$B = +0.3743936586512442 \times 10^{1}$
$C = -0.2769076432349482 \times 10^{1}$
$D = +0.9429681327692907 \times 10^{1}$
$E = +0.1055053194443676 \times 10^{2}$
$F = +0.6284207351324604 \times 10^{-2}$

$M = 0.9935108291614367 \times 10^{-13}$

$h = 4$

$A = +0.4507019590625572 \times 10^{0}$
$B = +0.3740448131455307 \times 10^{1}$
$C = -0.2768317205414987 \times 10^{1}$
$D = +0.9414899439055362 \times 10^{1}$
$E = +0.1053701311149719 \times 10^{2}$
$F = +0.6272339359526764 \times 10^{-2}$

$M = 0.1224323420423443 \times 10^{-10}$

$h = 16$

$A = +0.4405766736959988 \times 10^{0}$
$B = +0.3669989101432331 \times 10^{1}$
$C = -0.2752824996097087 \times 10^{1}$
$D = +0.9117484138879304 \times 10^{1}$
$E = +0.1026464040151929 \times 10^{2}$
$F = +0.6161613003116790 \times 10^{-2}$

$M = 0.2015977389469012 \times 10^{-6}$

$$S(X) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + a_6 x^6 \qquad \text{on } [-h, h]$$

$h = 1$

$$\begin{aligned}
a_0 &= +0.1666666666666665 \times 10^{0} \\
a_1 &= -0.8333333333333568 \times 10^{-2} \\
a_2 &= +0.1984126984129264 \times 10^{-3} \\
a_3 &= -0.2755731919939401 \times 10^{-5} \\
a_4 &= +0.2505210785999854 \times 10^{-7} \\
a_5 &= -0.1605953765319026 \times 10^{-9} \\
a_6 &= +0.7650283228592385 \times 10^{-12}
\end{aligned}$$

$M_a = 0.55511151 \times 10^{-16}$

$h = 2$

$$\begin{aligned}
a_0 &= +0.1666666666666663 \times 10^{0} \\
a_1 &= -0.8333333333352921 \times 10^{-2} \\
a_2 &= +0.1984126984129057 \times 10^{-3} \\
a_3 &= -0.2755731883077317 \times 10^{-5} \\
a_4 &= +0.2505210816170333 \times 10^{-7} \\
a_5 &= -0.1606101133600677 \times 10^{-9} \\
a_6 &= +0.7647926042737674 \times 10^{-12}
\end{aligned}$$

$M_a = 0.56621374 \times 10^{-14}$

$h = 4$

$$\begin{aligned}
a_0 &= +0.1666666666666581 \times 10^{0} \\
a_1 &= -0.8333333334593103 \times 10^{-2} \\
a_2 &= +0.1984126984258407 \times 10^{-3} \\
a_3 &= -0.2755731292513204 \times 10^{-5} \\
a_4 &= +0.2505210496761327 \times 10^{-7} \\
a_5 &= -0.1606691704048320 \times 10^{-9} \\
a_6 &= +0.7650122280184766 \times 10^{-12}
\end{aligned}$$

$M_a = 0.72000739 \times 10^{-12}$

$h = 16$

$$\begin{aligned}
a_0 &= +0.1666666661133027 \times 10^{0} \\
a_1 &= -0.8333338509758059 \times 10^{-2} \\
a_2 &= +0.1984127524406629 \times 10^{-3} \\
a_3 &= -0.2755570214946503 \times 10^{-5} \\
a_4 &= +0.2505123071826789 \times 10^{-7} \\
a_5 &= -0.1618528418504030 \times 10^{-9} \\
a_6 &= +0.7694603615375217 \times 10^{-12}
\end{aligned}$$

$M_a = 0.11845921 \times 10^{-7}$

$$q_1 = \sqrt[6]{a_6}\, x$$
$$q_2 = (q_1 + A)^2$$
$$q_3 = (q_2 + B)(q_1 + C)$$
$$S(X) = (q_2 + q_3 + D)(+q_3 + E) + F \quad \text{on } [-h, h]$$

$h = 1$

$A = +0.1030541110544949 \times 10^{0}$
$B = +0.1357446199107850 \times 10^{1}$
$C = -0.1709888012144409 \times 10^{1}$
$D = +0.1803960616654583 \times 10^{1}$
$E = +0.2062171852872241 \times 10^{1}$
$F = +0.2130010466488949 \times 10^{-1}$

$M = 0.2359223927328455 \times 10^{-14}$

$h = 2$

$A = +0.1028274494783523 \times 10^{0}$
$B = +0.1356879505417743 \times 10^{1}$
$C = -0.1709784631095070 \times 10^{1}$
$D = +0.1802832347589989 \times 10^{1}$
$E = +0.2060949727430426 \times 10^{1}$
$F = +0.2128753997322974 \times 10^{-1}$

$M = 0.8076872504148012 \times 10^{-14}$

$h = 4$

$A = +0.1026453527945748 \times 10^{0}$
$B = +0.1356011711587295 \times 10^{1}$
$C = -0.1709549340846876 \times 10^{1}$
$D = +0.1800557277079013 \times 10^{1}$
$E = +0.2059246874696585 \times 10^{1}$
$F = +0.2125209137884825 \times 10^{-1}$

$M = 0.7223804887601657 \times 10^{-12}$

$h = 16$

$A = +0.9898746283297480 \times 10^{-1}$
$B = +0.1338586121335949 \times 10^{1}$
$C = -0.1704756186376375 \times 10^{1}$
$D = +0.1754906650979839 \times 10^{1}$
$E = +0.2025050653157090 \times 10^{1}$
$F = +0.2056593593968585 \times 10^{-1}$

$M = 0.1184592103575797 \times 10^{-7}$

$$C'(X) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + a_6 x^6 \text{ on } [-h, h]$$

$h = 1$

$a_0 = -0.4166666666666430 \times 10^{-1}$
$a_1 = +0.2777777778467318 \times 10^{-2}$
$a_2 = -0.1488095238678453 \times 10^{-3}$
$a_3 = +0.6613751098214413 \times 10^{-5}$
$a_4 = -0.2505208412532178 \times 10^{-6}$
$a_5 = +0.8269965205281566 \times 10^{-8}$
$a_6 = -0.2412214891589441 \times 10^{-9}$

$M_a = 0.21684043 \times 10^{-16}$

$h = 2$

$a_0 = -0.4166666666606742 \times 10^{-1}$
$a_1 = +0.2777777822034584 \times 10^{-2}$
$a_2 = -0.1488095275535492 \times 10^{-3}$
$a_3 = +0.6613668137463213 \times 10^{-5}$
$a_4 = -0.2505171918626241 \times 10^{-6}$
$a_5 = +0.8303136595603477 \times 10^{-8}$
$a_6 = -0.2422316681583509 \times 10^{-9}$

$M_a = 0.25058081 \times 10^{-14}$

$h = 4$

$a_0 = -0.4166666651125364 \times 10^{-1}$
$a_1 = +0.2777780643726382 \times 10^{-2}$
$a_2 = -0.1488097663598733 \times 10^{-3}$
$a_3 = +0.6612326049924104 \times 10^{-5}$
$a_4 = -0.2504581368842600 \times 10^{-6}$
$a_5 = +0.8437138609417991 \times 10^{-8}$
$a_6 = -0.2463133603265637 \times 10^{-9}$

$M_a = 0.31997321 \times 10^{-12}$

$h = 16$

$a_0 = -0.4166666641763858 \times 10^{-1}$
$a_1 = +0.2777780078584500 \times 10^{-2}$
$a_2 = -0.7440478621862503 \times 10^{-4}$
$a_3 = +0.1102220894089232 \times 10^{-5}$
$a_4 = -0.1043798352532477 \times 10^{-7}$
$a_5 = +0.6938557071056230 \times 10^{-10}$
$a_6 = -0.3366982823031963 \times 10^{-12}$

$M_a = 0.52654049 \times 10^{-8}$

$$q_1 = \sqrt[6]{-a_6}\ x$$

$$q_2 = (q_1 + A)^2$$

$$q_3 = (q_2 + B)(q_1 + C)$$

$$C'(X) = (q_2 + q_3 + D)(-q_3 - E) - F \quad \text{on } [-h, h]$$

$h = 1$

$A = +0.6421500675794880 \times 10^{-1}$
$B = +0.9631412054424315 \times 10^{0}$
$C = -0.1485378935681960 \times 10^{1}$
$D = +0.1206516480446427 \times 10^{1}$
$E = +0.1262370275488431 \times 10^{1}$
$F = +0.2235785774036898 \times 10^{-2}$

$M = 0.1786765180256109 \times 10^{-15}$

$h = 2$

$A = +0.6415596248585610 \times 10^{-1}$
$B = +0.9629599640982438 \times 10^{0}$
$C = -0.1485350658374242 \times 10^{1}$
$D = +0.1206153148806730 \times 10^{1}$
$E = +0.1262088821410514 \times 10^{1}$
$F = +0.2230745797490601 \times 10^{-2}$

$M = 0.2680147770384164 \times 10^{-14}$

$h = 4$

$A = +0.6405925631551870 \times 10^{-1}$
$B = +0.9624584033933687 \times 10^{0}$
$C = -0.1485269967153406 \times 10^{1}$
$D = +0.1205018912730453 \times 10^{1}$
$E = +0.1261394717119364 \times 10^{1}$
$F = +0.2210887038704210 \times 10^{-2}$

$M = 0.3200469403386028 \times 10^{-12}$

$h = 16$

$A = +0.6208183431484613 \times 10^{-1}$
$B = +0.9523193847619871 \times 10^{0}$
$C = -0.1483582934430130 \times 10^{1}$
$D = +0.1182131124018069 \times 10^{1}$
$E = +0.1247277466997771 \times 10^{1}$
$F = +0.1829570481293727 \times 10^{-2}$

$M = 0.5265405070287163 \times 10^{-8}$

$$S'(X) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + a_6 x^6 \qquad \text{on } \left[-h,\ h\right]$$

$h = 1$

$a_0 = -0.8333333333333210 \times 10^{-2}$
$a_1 = +0.3968253968616768 \times 10^{-3}$
$a_2 = -0.1653439153716376 \times 10^{-4}$
$a_3 = +0.6012503110063301 \times 10^{-6}$
$a_4 = -0.1927084107177350 \times 10^{-7}$
$a_5 = +0.5511761621292874 \times 10^{-9}$
$a_6 = -0.1418571972378231 \times 10^{-10}$
$M_a = 0.26020852 \times 10^{-17}$

$h = 2$

$a_0 = -0.8333333333304809 \times 10^{-2}$
$a_1 = +0.3968253991531185 \times 10^{-3}$
$a_2 = -0.1653439171256403 \times 10^{-4}$
$a_3 = +0.6012459474285134 \times 10^{-6}$
$a_4 = -0.1927066737597982 \times 10^{-7}$
$a_5 = +0.5529210296030711 \times 10^{-9}$
$a_6 = -0.1423381311296310 \times 10^{-10}$
$M_a = 0.13444107 \times 10^{-15}$

$h = 4$

$a_0 = -0.8333333325953303 \times 10^{-2}$
$a_1 = +0.3968255472561213 \times 10^{-3}$
$a_2 = -0.1653440305444520 \times 10^{-4}$
$a_3 = +0.6011754909568150 \times 10^{-6}$
$a_4 = -0.1926786201990815 \times 10^{-7}$
$a_5 = +0.5599576811597955 \times 10^{-9}$
$a_6 = -0.1442776137237285 \times 10^{-10}$
$M_a = 0.16841563 \times 10^{-13}$

$h = 16$

$a_0 = -0.8333333321480760 \times 10^{-2}$
$a_1 = +0.3968255178485000 \times 10^{-3}$
$a_2 = -0.8267196924462379 \times 10^{-5}$
$a_3 = +0.1002046526767437 \times 10^{-6}$
$a_4 = -0.8029333915166488 \times 10^{-9}$
$a_5 = +0.4617817733427159 \times 10^{-11}$
$a_6 = -0.1978183008506138 \times 10^{-13}$
$M_a = 0.27690121 \times 10^{-9}$

$$q_1 = \sqrt[6]{-a_6}\ x$$

$$q_2 = (q_1\ A)^2$$

$$q_3 = (q_2 + B)\ (q_1 + C)$$

$$S'(X) = (q_2 + q_3 + D)\ (-q_3 - E) - F \quad \text{on } [-h,\ h]$$

$h = 1$

$A = -0.4272502910304863 \times 10^{-1}$
$B = +0.4460409572215307 \times 10^{0}$
$C = -0.1020276052465536 \times 10^{1}$
$D = +0.4064530281034859 \times 10^{0}$
$E = +0.3450015283629582 \times 10^{0}$
$F = +0.2885054406289098 \times 10^{-2}$

$M = 0.5290906601729259 \times 10^{-16}$

$h = 2$

$A = -0.4272336986330691 \times 10^{-1}$
$B = +0.4460090638507059 \times 10^{0}$
$C = -0.1020282840851638 \times 10^{1}$
$D = +0.4064007086884238 \times 10^{0}$
$E = +0.3449898120978949 \times 10^{0}$
$F = +0.2883373149053849 \times 10^{-2}$

$M = 0.1821459649775645 \times 10^{-15}$

$h = 4$

$A = -0.4274056995233864 \times 10^{-1}$
$B = +0.4458621272744412 \times 10^{0}$
$C = -0.1020307887595762 \times 10^{1}$
$D = +0.4061739766937458 \times 10^{0}$
$E = +0.3449170623874602 \times 10^{0}$
$F = +0.2876661053629565 \times 10^{-2}$

$M = 0.1689186984732414 \times 10^{-13}$

$h = 16$

$A = -0.4310245077956152 \times 10^{-1}$
$B = +0.4428994350437274 \times 10^{0}$
$C = -0.1020789569928175 \times 10^{1}$
$D = +0.4016042135773923 \times 10^{0}$
$E = +0.3434241762739658 \times 10^{0}$
$F = +0.2744481265299470 \times 10^{-2}$

$M = 0.2769012918263367 \times 10^{-9}$

$$P(X) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$$

$C(X)$ #4 $[-1, +1]$

$$a_0 = +0.5000000000007167 \times 10^0$$
$$a_1 = -0.4166666601424471 \times 10^{-1}$$
$$a_2 = +0.1388888879568642 \times 10^{-2}$$
$$a_3 = -0.2480419696201834 \times 10^{-4}$$
$$a_4 = +0.2755932664579228 \times 10^{-6}$$

$$M_a = 0.13048673 \times 10^{-9}$$

$S(X)$ #4 $[-1, +1]$

$$a_0 = +0.1666666666667142 \times 10^0$$
$$a_1 = -0.8333333283147393 \times 10^{-2}$$
$$a_2 = +0.1984126977913694 \times 10^{-3}$$
$$a_3 = -0.2755932664326337 \times 10^{-5}$$
$$a_4 = +0.2505344666229896 \times 10^{-7}$$

$$M_a = 0.10037290 \times 10^{-10}$$

$$q_1 = Ax$$
$$q_2 = (q_1 + B)^2$$
$$P(x) = (q_1 + q_2 + C)(q_2 + D) + E$$

$C(X)$ #4 $[-1, +1]$

$$A = +0.2291221893900772 \times 10^{-1}$$
$$B = -0.7655416156428518 \times 10^0$$
$$C = +0.2592589041826887 \times 10^0$$
$$D = +0.4011554686880556 \times 10^0$$
$$E = -0.3345008393894142 \times 10^0$$

$$M = 0.1304868574303339 \times 10^{-9}$$

$S(X)$ #4 $[-1, +1]$

$$A = +0.1258104947765253 \times 10^{-1}$$
$$B = -0.5959855810891701 \times 10^0$$
$$C = +0.1104585254341674 \times 10^0$$
$$D = +0.2038526154314646 \times 10^0$$
$$E = -0.9365973340739070 \times 10^{-1}$$

$$M = 0.1003741534333355 \times 10^{-10}$$

REFERENCES

1. Battin, Richard H.: Astronautical Guidance, McGraw Hill, Inc., 1964.

2. Stoer, Josef: A Direct Method for Chebyshev Approximation by Rational Functions. J. ACM, Vol. 11, No. 1 (January, 1964), pp. 59-69.

3. Fike, C. T.: Methods of Evaluating Polynomial Approximations in Function Evaluation Routines. Comm. ACM, Vol. 10, No. 3 (March, 1967), pp. 175-178.

4. Knuth, D. E.: Evaluation of Polynomials by Computer. Comm. ACM, Vol. 5, No. 12 (December, 1962), pp. 595-599.

5. See Todd, John: Motivation for Working in Numerical Analysis. Comm. Pure and Appl. Math., Vol. VIII (1955), pp. 98-100.

APPENDIX I

The cubic equation (16) was solved by an interval-halving technique which can be extended to any continuous function f. In general, this technique gives us the "smallest" (in the sense of representability by computer) interval in which a value x can lie such that $f(x) = 0$, and this smallest interval can be found by a finite, fixed number of iterations. The most familiar interval-halving process consists of two parts: (1) finding the initial bracketing interval, the interval $[a, b]$ in which $f(x)$ changes sign, then (2) successively halving this interval, choosing each time the subinterval in which $f(x)$ changes sign, until the interval is as small as desired. If part (1) is performed properly, then part (2) can be performed with a number of iterations determinable a priori, thus eliminating the test for interval size at the end of each iteration.

To begin, choose a number v and a number t, of the same sign as v, such that $|t| \leq |v|$. [Positive v's are used for positive roots and negative v's for negative roots.] Test the following sequence of intervals for a sign change of $f(x)$:

$$(1) \quad [v,\ v + t]$$

$$(2) \quad [v + t,\ v + t + 2t]$$

$$(3) \quad [v + t + 2t,\ v + t + 2t + 4t]$$

$$\vdots$$

$$(p + 2) \quad \left[v + \sum_{i=0}^{p} 2^i t,\ v + \sum_{i=0}^{p+1} 2^i t\right]$$

until the initial bracketing interval is found. Since $|t| \leq |v|$, we have

$$|v| + 2^{q+1}|t| - |t| \geq 2^{q+1}|t| \quad \Rightarrow$$

$$|v| + |t| \left(2^{q+1} - 1\right) \geq 2^{q+1}|t| \quad \Rightarrow$$

$$|v| + |t| \sum_{i=0}^{q} 2^i \geq |2^{q+1} t|$$

and since v and t have the same sign, we have

$$\left| v + \sum_{i=0}^{q} 2^i t \right| \geq |2^{q+1} \; t|$$

This simply says that the length of the initial bracketing interval is less than, or equal to, the magnitude of the small end; in turn, this means that the large end of the interval is at most twice the magnitude of the small end.

Now, consider how the endpoint values would be represented in floating-point binary arithmetic (normalized) with an r-bit fraction. If the difference between their binary exponents is at most 1 (which is what we are getting at above), then it can be seen that the number of distinct points in the initial bracketing interval is at most 2^r. Therefore, the number of interval-halving iterations needed—that is, the number of times one reduces his choice of points in the interval by one-half—is at most r. Moreover, it often turns out that f is nearly (or exactly) zero at an end point of one of the half-intervals, so that r iterations are not always needed.

We have treated the special case $|t| \leq |v|$, but we need not restrict ourselves to it. The number of interval-halving iterations needed depends upon the size of t, and if one is willing to iterate a bit more he can find the initial bracketing interval more quickly by increasing t; the converse of this also holds.